Eric Hu

Dr. Anna Lesne

Treatment and Analysis with Advertisement Data

May 30, 2023

Description: This research stems from a group project at the ACSS Institute - PSL that is oriented towards ARCOM. I will be mainly focusing on my role in the team—managing and analyzing the three-year French advertisement data provided by ARCOM. There are two main themes in the interests of data exploration: sexism and greenwashing. To detect signs and trends of these themes, one should not only be familiar with the rhetorical features but more importantly the relevant data science techniques. Thus, I will build an overarching framework and illustrate the methodology needed in the analyzing process based on my internship experiences and observations from academic sources.

Exploratory Analysis

When we start the analysis of a dataset, one of the first and foremost actions that we shall take is exploratory analysis. One may go in any direction, such as data type, size, and frequency, so that we obtain an overall understanding of the dataset.

a. Identifying useful features

Since there are multiple features, in this ARCOM case, 37 columns in the dataset 'ARCOM_scripts_2020_2022.xlsx' (ARCOM, 2022), it is indispensable to select and focus on a few key features. The first column that I examined in the ARCOM dataset is "N° Fiche" (File Number) which should be a unique identifier for every documented advertisement. However, I discovered some discrepancies via the command that shows the counting of all values in the data frame: df['N° Fiche'].value_counts(dropna=False). Because each file number is unique, their count in the result should be exactly equal to 1. Nevertheless, there were hundreds of counts equal to 2, indicating the existence of duplicate entries. Thus, I located those file numbers shown in the previous result and confirmed that the advertisement data for Week 31 and 32 are exactly identical, and the same case for Week 36 and 37. The way used to confirm duplicates is to check the file size within the raw data, which includes Excel sheets that each stand for one week in a year. For Week 31/32 and 36/37, they have exactly the same file size, which is impossible unless the two data files are identical in every character. Finally, I reported my discovery to the team coordinator in the second week of my internship. This became my first contribution to the team and the coordinator, also my supervisor, succeeded to obtain a corrected version from ARCOM that allowed us to conduct the following analysis.

Furthermore, the team received a dictionary from ARCOM that explains the usage of each column. We then discussed and decided to focus on the following 8 features: “Annonceur” which specifies the advertiser; “Produit” which denotes the name of the product in the Ad; “Secteur–Classe–Groupe–Variété”, the four features that classify the Ad from broad sector to detailed variety; “Script”, the most important feature that records the speakers and their conversations; “Mots Clés”, keywords that summarise the content in the Ads. The next step would be digging into these crucial features.

b. Performing basic statistics

Statistics is one of the simplest ways to obtain an overview of the 43286 Ads and generate some meaningful numbers. I was in charge of computing the distribution of different values within Section/Advertiser/Product. The command used is almost the same as the last one:

df['Secteur'].value_counts(dropna=False, normalize=True).mul(100).round(1).astype(str) + '%'. Instead of giving the counts of each value, the “normalize” feature divides the counts by the total counts so that by multiplying 100 we could get the percentages of each unique value within a specific field. For example, these are the top 5 sectors’ occurrences in the “Section” column:

```
ALIMENTATION              17.4%
CULTURE & LOISIRS         10.8%
DISTRIBUTION               8.8%
TOURISME-RESTAURATION   7.0%
SERVICES                   6.3%
```

One could clearly see that the most frequent section in the dataset is “Alimentation (Food)”. These figures are quite flexible in their usage. My colleague adopted them to make three bar charts “Distribution of the 10 most frequent Sectors/Advertisers,” as shown in Figure A1/2 in Appendix (Fumeron, 2023). This method is also useful in displaying the result, which would be used again in the following Gender Prediction.

Data Cleaning

Before conducting any textual analysis, we need to understand what the “text” is from a data science perspective. Although we may have multiple definitions, “for our purposes, we will think of a text as nothing more than a sequence of words and punctuation. ” (Bird, 2009, p. 10). Words and punctuations, as mentioned by Steven Bird, constitute the object of our textual analyzing work. In this section, I will discuss several data-cleaning techniques that help to make the text more usable before applying actual natural language models. According to Edward Loper, “Raw data needs to be collected, cleaned up, documented, and stored in a systematic structure.” (Loper, 2009, p. 412). Nevertheless, data cleaning is often neglected in terms of its significance. There is an exaggerated saying among my data scientists colleagues that 80% of the work time should be assigned to cleaning the data and only 20% used to adjust and apply the models on the cleaned data. Hence, the importance of cleaning data in a delicate way couldn’t be more stressed. In the following parts, I will discuss the approaches taken by my team to clean up the texts.

a. Separating narrators and quotes

To study the gender occurrences in an Ad, we focused entirely on the “Script” feature since it is the only column that contains information about the speaker. For instance, in the Ad by CHERIE FM that has celebrities wishing a happy new year, we have the following script (ARCOM, 2022):

> Angèle (1) : " Bonne année sur CHERIE FM. "
> Christophe Maé (2) : " Bonne année sur CHERIE FM.
> M. Pokora (3) : " Bonne année 2020. "
> Vitaa & Slimane (4) : " CHERIE FM. " (2) ... Le plus belle musique. "
> Clara Luciani (5) : " Bonne année sur CHERIE FM. "
> Pink (6) : " Bonne année, ... "
> Sam Smith (7) : " CHERIE FM.

The script is conducted in the format of Role: (No.) “Quotes.” Based on this format, we came up with an algorithm to separate the narrators and their sentences. This algorithm was designed by my colleague and the key is to utilize the concept of “Regular Expression” in computer science. “Regular expressions are combinations of special character operators, which are symbols that control the search, that you can use to construct search strings (words) for advanced find and/or replace searches.” (Intel, n.d.). In the example provided by Steven Bird, we are able to match the word “email” and “e-mail” using the regular expression “^e-?mail$,” where “the caret symbol ‘^’ matches the start of a string…the ‘$’ matches the end…the ‘?’ symbol specifies that the previous character is optional.” (Bird, 2009, p .99). Since regular expressions make the computer go through the dataset and look for a specific pattern, we used the regular expression "^\s*\(.*\)\s*$" to match the pattern made up of punctuations (): “ which separates the role and quote as shown in the ARCOM script. Then, we designed another algorithm to catch the part before the pattern as a role name and the part after it as a quote spoken by the role. This step is crucial since role names and quotes are different entities in terms of ontology. We would like to study the characters' names and quotes individually. In the Data Manipulation Section, I would make gender predictions based on the role names using the cleaned data obtained here.

b. Removing punctuation and stopwords

Before applying most models, it's essential to remove anything unrelated to the content as much as possible. Not only does it simplify the text, but it also improves the results of most algorithms. To begin with, we shall remove punctuation from the sentences because they contain little semantics. The way through which we managed to remove all punctuations is easy since there is a built-in command in Python, following an example given by Steven Bird: "To derive the vocabulary, collapsing case distinctions and ignoring punctuation, we can write set([w.lower() for w in text if w.isalpha()])." (Bird, 2009, p. 33).

Except for punctuation, we also need to take care of stopwords, defined as a set of commonly used words in any language. In English, common stopwords include "the", "and", etc. In the French language that we dealt with, there was an entire list of stopwords, such as "ça", "être", and "un". These words hardly help in determining the meaning of a large paragraph. We were looking for words with strong meanings like "nature" or "polluter". Hence, it is essential to specify the list of stopwords before we perform the analysis so that the computer ignores these frequent yet barely useful words. In the majority of language processing models, there is an option for "stopwords" and simply feeding it with the premade stopword list would have the models automatically neglect these stopwords.

Through this step, we managed to reduce the noise significantly in our data and make it suitable for further analysis.

c. Making up for missing values

When there are missing key values that we are supposed to study in the dataset, we often have two options: dropping these entries (i.e., abandoning the entire row if one of the essential values, such as Script, is missing) or maintaining the entries with a designed substitute method. The strategy that my team took against the dataset is as follows: if we missed almost all useful features identified in the previous section, we should consider dropping these values since none of the techniques introduced can be applied. However, as long as we can make up for the missing values, for instance, some Ad entries with scripts but without keywords, we should keep those entries. I came up with a way to save those entries with "Script" but without "Mots Clés." by extracting keywords from their scripts and using those as the initially missing keywords. To put it simply, I utilized the existing features to infer those missing features whenever possible. I browsed many keyword extractors with their documentation on the internet and decided to use KeyBert, which is slightly different from other extraction techniques adopted by my colleagues—we ran different techniques and compared the results. I will talk about the keyword extraction method more detailedly in Section Data Manipulation since it is extremely useful when we conduct thematic research, especially for the sake of topic modeling.

Data Manipulation

As Steven Bird points out in algorithm design, "a major part of algorithmic problem solving is selecting or adapting an appropriate algorithm for the problem at hand." (Bird, 2009, p. 160). Before choosing an appropriate algorithm, we need a good understanding of the problem that we need to solve. Hence, in the next section, I will first cover the tasks and then introduce the optimal solutions to these tasks that the team has identified.

a. Gender prediction via gender-guesser and complementary methods

In the previous section, we obtained a list of role names and quotes from the Script. This form of data is defined to be "*structured data*, where there is a regular and predictable organization of entities and relationships." (Loper, 2009, p. 261). To study sexism in French Ads, the team decided to study the number of males and females that appeared in Ads in the three-year period. I was in charge of designing the prediction algorithm with the structured data provided. Given an existing Python package called "gender-guesser" that makes predictions based on the first names in Francophone areas (Elmas, 2021), I added my extra version of the classifier into the algorithm to greatly improve its prediction accuracy. When the gender-guesser has trouble with roles without first names in the Ads, for instance, "Voix Homme," I manually created a list that contains almost all ways of addressing a man {Mr., homme, garçon, fils...} (the same case for females). As long as the role name contains any word in the male/female addressing list, the model is set to predict "male/female." Thus, this algorithm became more robust: if the role name contains male/female descriptions, it will directly output its prediction; if the role name contains a first name and last name, it will input its first name into the gender guesser and output the result from the guesser; if neither case works, it outputs "unidentifiable," for instance, when the role name is "Voix enfants." This algorithm is another of my contributions to the ACSS team, and it was presented to ARCOM during a routine meeting.

With the gender prediction for every role in the whole dataset, we were capable of calculating the percentage of men and women speakers in each Ad sector. There are results that conform to our expectations: in Sector "Hygiène Beauté," there are 69% female roles in the Ads as opposed to 21% male roles. However, there are also unexpected findings: in Sector "Batiment Travaux Publics," there are 63% females and only 33% males. My colleague visualized the results for each sector with Graph B1/2 in the Appendix. (Fumeron, 2023). Currently, we only have a limited amount of data (2020–2023). Nevertheless, this model will be of greater significance given a broader range of data; for instance, we can clearly see how the male/female percentage has evolved over decades if we have enough annual advertisement data.

b. Extracting Keywords

“Keywords and keyphrases play an important role in getting the idea behind text data quickly without having to read through the whole text.” (Sharma & Li, 2019, p. 1). As stated by Sharma and Li, in order to study the conversations in all 43286 Ads, one of the most efficient methods is to extract keywords/keyphrases from the bulk scripts. “The keywords and keyphrases retrieval methods can be broadly classified into following: statistical approach, graphbased approach, linguistic approach, machine learning approach and hybrid approach.” (Sharma & Li, 2019, p. 2). At the moment, the hybrid approach is the most popular since it combines the advantages of all other approaches. My colleagues ran some traditional keyword extraction techniques like TF-IDF. They are based upon the statistical properties of the texts, the principles of which stem from calculating the TF-IDF of a word inside the whole document. Unlike those traditional methods, KeyBert takes into account the semantic aspects of the whole document and produced more reliable results that were extremely useful when we conducted thematic research, especially in identifying ads that were related to greenwashing. Here is a summarized description of the algorithm behind it: First, the whole document is fed to the BERT model to get a document-level representation (i.e., a vector, which one may understand as a quantity having direction as well as magnitude, or the coordinates in a multi-dimensional space). Then, each word goes through the same procedure that results in word-level representation. Finally, we use cosine similarity to find the words/phrases that are the most similar to the document.

(Grootendorst, 2021). Hence, the most similar words could then be identified as the words that best describe the entire document. For example, in the following Ad script without original keywords, we may obtain via KeyBert a list of keywords [cérébral, 2020, kawashima, duel, switch] from the following text (ARCOM, 2022):

> "En 2020, quel âge à votre cerveau ?... Exercice après exercice ...... améliorez votre âge cérébral.... Affrontez-vous même à deux dans un duel de matières grises.... Le programme d'entraînement cérébral ...... du docteur Kawashima pour NINTENDO SWITCH, votre nouvelle résolution. "

With these machine-generated keywords, we are able to, as mentioned earlier, make up for the entries without “Mots Clès (keywords)” and even compare them to those with original “keywords.” Since the original keywords are not entirely based on the Script, (there are other important aspects like visual/audio effects), their comparison with script-based keywords could elicit some interesting discoveries.

c. Zero-shot Classification

In the last section, given the original script, we successfully produced a list of labels most relevant to the text. Now, we consider the converse: given a list of labels, to what extent is it relevant to the script? Through zero-shot classification, we are able to determine the relevance of a premade word to any type of text. “Zero-shot learning (ZSL) most often referred to a fairly specific type of task: learn a classifier on one set of labels and then evaluate on a different set of labels that the classifier has never seen before.” (Davison, 2020). The classification process is similar to how KeyBert extraction works (they are two sides of one problem): if we provide the pre-trained model with a new label, it is able to predict the similarity between the new label and the text based on their representation vector. With zero-shot classification, we figured out a way to screen out those potential greenwashers. Notice that sector labels aren’t as helpful in classifying greenwashers which could exist in several sectors—Food, Appliance, Construction, etc. For example, I chose some common environment-related words to form a label list: ['Nature', 'Ecologie', 'Environnement', 'Green', 'Biodiversité', 'Climatique', 'Développement durable', 'Energies renouvelables']. Then, I called zero-shot classification upon the following Ad of Lipton: (ARCOM, 2022)

> "Bienvenue chez LIPTON, ... dans notre usine de thé.... Ici, nous laissons faire la nature.... Voici notre système d'irrigation, ... la pluie.... Et là, ... notre source de lumière, le soleil ...... qui fait grandir nos feuilles dans les meilleures conditions.... C'est cette nature, alliée au savoir-faire LITPON, ... qui nous permet de vous proposer des thés de grande qualité.... Noté excellent sur Yuka.... LITPON.... Découvrez aussi la gamme LITPON BIO. "

The relevance score for each word in the list is [0.66, 0.16, 0.06, 0.05, 0.04, 0.01, 0.006, 0.004] on a scale from 0 to 1. We found that the word ‘Nature’ as a label has the highest relevance score 0.66, which points out a strong relationship between the word and the text. Because any label in the list above having a high relevance score against the script could suggest potential greenwashing, I proposed a strategy that the highest relevance score across the label list computed by zero-shot classification can become an effective indicator of eco-related Ads, which would help us detect Ads with a high probability of being greenwashers. So far, we are unable to know whether the Ad is a greenwasher simply by confirming its eco-theme. This would require manual inscription or a much more sophisticated algorithm to achieve. To use the words of K.R.Chowdhary, “Automatic analysis of text requires a deep understanding of natural language by machines. However, we are still far away from machines that have this capability…” (Chowdhary, 2020, p. 645). At the moment, ChatGPT might have the potential to crack this problem.

d. Topic Modeling

Topic modeling is a major field in natural language processing that uses unsupervised machine learning techniques. The key difference between unsupervised learning and supervised learning (adopted by the previous models) is that “supervised learning uses labeled datasets, whereas unsupervised learning uses unlabeled datasets." (Alteryx, n.d.). According to Kherwa and Bansal, "topic modeling is a technique [that] comes with [a] group of algorithms that reveal, discover and annotate thematic structure in [a] collection of documents" (Kherwa & Bansal, 2018, p. 2).

Given a paragraph of texts, the topic modeling algorithm will generate different clusters of similar words, providing us with potential topics of very large textual data in an efficient way. The machine only groups words that it deems to be similar together, without any previous knowledge, or given labels, of the text. The way to achieve this is, again, through representation. Through representation from words to vectors, the proximity in the vector space implies proximity in the meanings of the words. The transformer model eventually returns the set of vectors close in mathematical space and the words that they represent as a cluster to form a topic. My team adopted topic modeling based on keywords grouped by different sectors in the dataset. The results produced 8 different topics within a certain sector. For instance, in "Culture & Loisirs," there is a topic constituted with these words [harry, potter, magique, barbie] and another topic [concert, paris, france, jazz]. We can navigate through a specific sector, which is huge in the number of Ads, by studying the topics graph generated via topic modeling, as shown in Figure C1/2 in the Appendix (Carlos, 2023). One may notice that there are quite a few spelling mistakes in the dataset. Except for the fact that the dataset is created manually by personnel at ARCOM, it has revealed several places to improve while preparing the data and choosing the model. Since my colleague utilized a topic modeling model that automatically removes accents in words, it resulted in several misspelled words in the topics. Besides, we could use lemmatization to further "clean" the data before running the model. For example, after lemmatization, "better" will become "good" and "eating/eaten" will become "eat." This grants words more weight by reducing all their different conjugations, which leads to less repetition in the final result. It is a common function in most natural language processing models.

Conclusion: I would like to thank all my colleagues at ACSS and Paris Dauphine University for this amazing experience working with data scientists. During this internship, I have learned much about data analysis, especially about natural language processing. I wrote this paper not only to record the common data-analyzing techniques but also to discuss the principles behind them. I hope this can help someone learn more about the internship at ACSS -PSL.

Reference

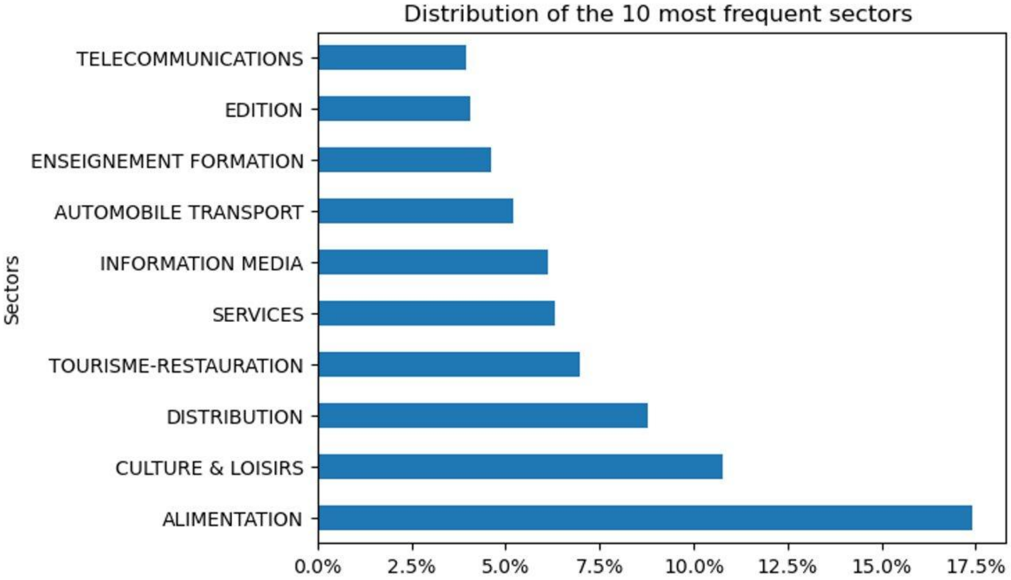
Primary sources:

1. ARCOM. (2022). *ARCOM_scripts_2020_2022.xlsx* [Internal dataset]. ACSS - PSL, France
2. Fumeron. P. & Carlos. J. (2023). *Presentation ARCOM* [Internal PowerPoint presentation]. ACSS - PSL, France
3. ACSS-PSL. (2023). *ARCOM* [Github repository]. https://github.com/ACSS-PSL/Arcom
4. Intel. (n.d.). *Regular Expression Definition.* Quartus Help. https://www.intel.com/content/www/us/en/programmable/quartushelp/17.0/reference/glossary/def_reg_express.htm
5. Elmas, F. (2021). *Gender Guesser.* GitHub repository. https://github.com/lead-ratings/gender-guesser
6. Grootendorst, M. (2021). *KeyBERT.* GitHub repository. https://github.com/MaartenGr/KeyBERT
7. Davison, J. (2020, May 29). *Zero-Shot Learning with CLIP and StyleGAN.* https://joeddav.github.io/blog/2020/05/29/ZSL.html
8. Alteryx. (n.d.). *Supervised vs. unsupervised learning.* https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning#:~:text=Supervised%20and%20unsupervised%20learning%20have,tagged%20with%20the%20right%20answer.&text=A%20classification%20problem%20uses%20algorithms%20to%20classify%20data%20into%20particular%20segments.

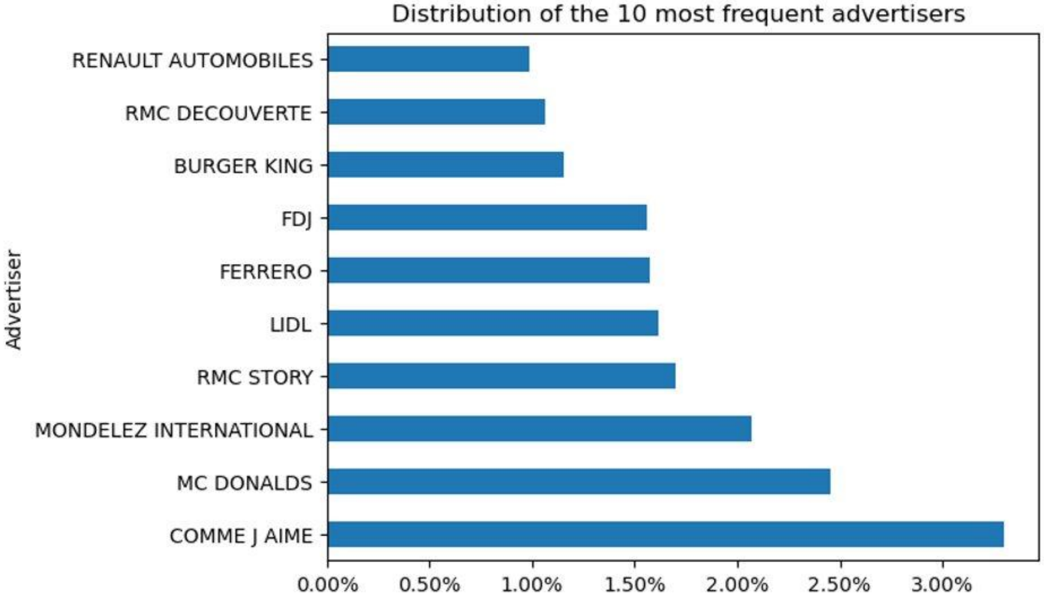
Scholarly Sources:

1. Bird, S., Klein, E., & Loper, E. (2009). Chapter 1: Language Processing and Python. *In Natural Language Processing with Python.* O'Reilly Media. https://tjzhifei.github.io/resources/NLTK.pdf
2. Bird, S., Klein, E., & Loper, E. (2009). Chapter 3: Processing Raw Text. *In Natural Language Processing with Python.* O'Reilly Media.
3. Bird, S., Klein, E., & Loper, E. (2009). Chapter 4: Writing Structured Programs. *In Natural Language Processing with Python.* O'Reilly Media.
4. Bird, S., Klein, E., & Loper, E. (2009). Chapter 7: Extracting Information from Text. *In Natural Language Processing with Python.* O'Reilly Media.
5. Bird, S., Klein, E., & Loper, E. (2009). Chapter 11: Managing Linguistic Data. In Natural Language Processing with Python. O'Reilly Media.
6. Sharma, P., & Li, Y. (2019). *Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling.* Published online on https://www.preprints.org/manuscript/201908.0073/v1
7. Chowdhary, K.R. (2020). Chapter 19: Natural Language Processing. *Natural Language Processing. In: Fundamentals of Artificial Intelligence.* Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19
8. Kherwa, P. & Bansal, P. (2018). *Topic Modeling: A Comprehensive Review.* ICST Transactions on Scalable Information Systems. 7. 159623. doi:10.4108/eai.13-7-2018.159623.

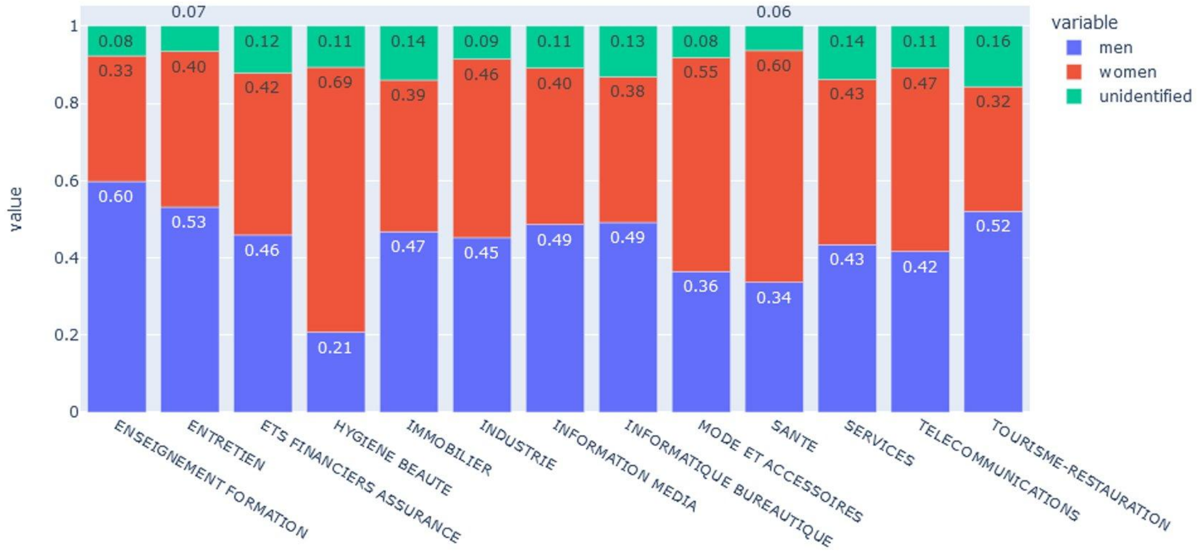
Appendix



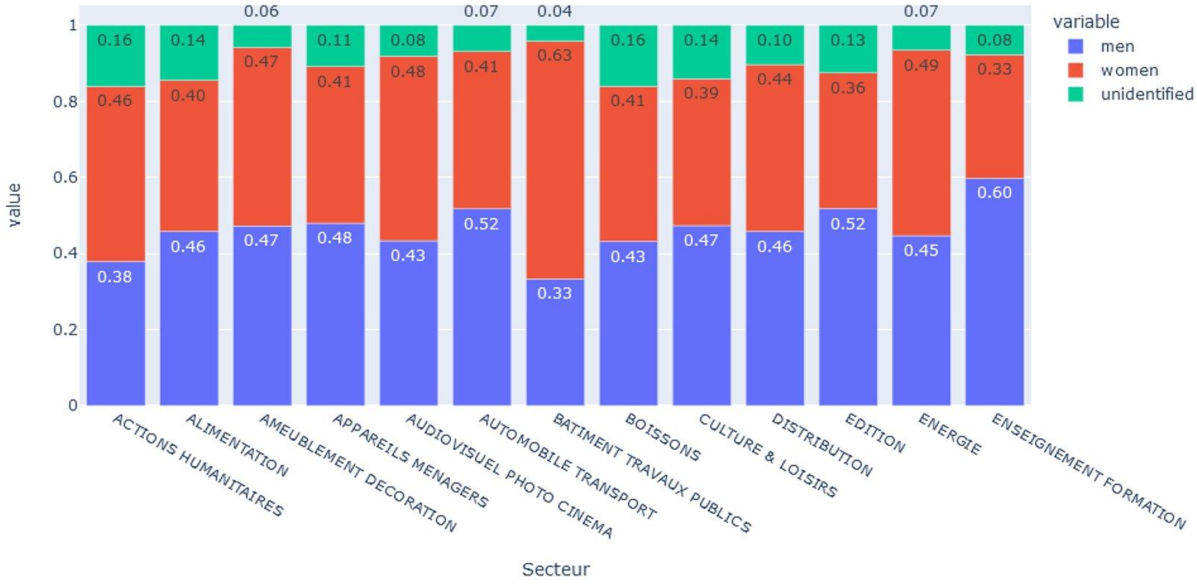
A1



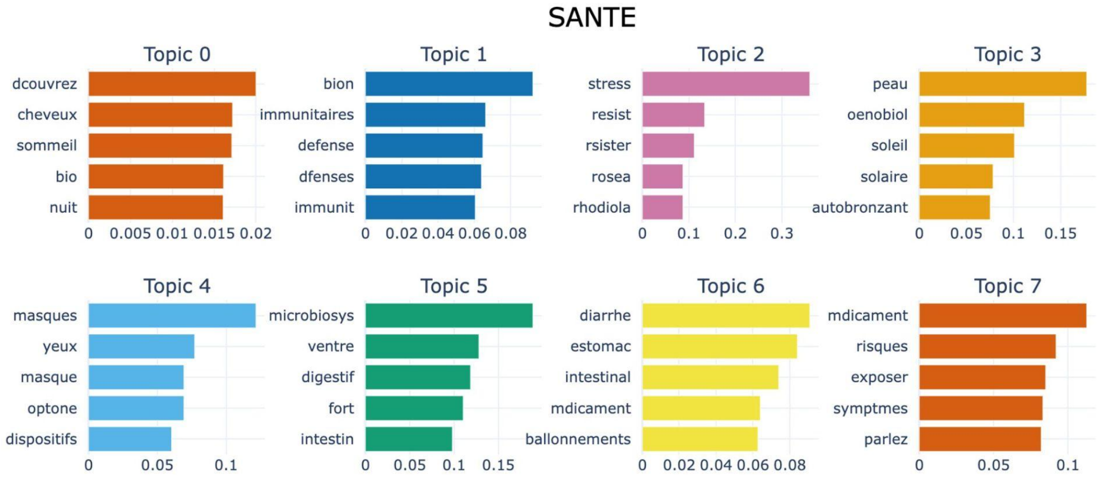
A2


variable
🟦 men
🟥 women
🟩 unidentified
value
0
0.2
0.4
0.6
0.8
1
0.08
0.33
0.60
0.07
0.40
0.53
0.12
0.42
0.46
0.11
0.69
0.21
0.14
0.39
0.47
0.09
0.46
0.45
0.11
0.40
0.49
0.13
0.38
0.49
0.08
0.55
0.36
0.06
0.60
0.34
0.14
0.43
0.43
0.11
0.47
0.42
0.16
0.32
0.52
ENSEIGNEMENT FORMATION
ENTRETIEN
ETS FINANCIERS ASSURANCE
HYGIENE BEAUTE
IMMOBILIER
INDUSTRIE
INFORMATION MEDIA
INFORMATIQUE BUREAUTIQUE
MODE ET ACCESSOIRES
SANTE
SERVICES
TELECOMMUNICATIONS
TOURISME-RESTAURATION

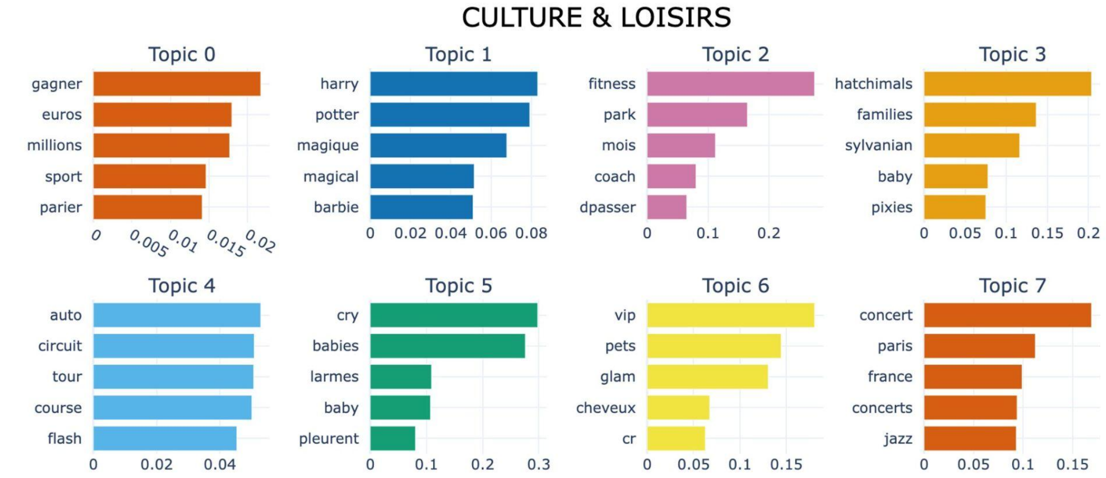
B1


variable
🟦 men
🟥 women
🟩 unidentified
value
0
0.2
0.4
0.6
0.8
1
0.16
0.46
0.38
0.14
0.40
0.46
0.06
0.47
0.47
0.11
0.41
0.48
0.08
0.48
0.43
0.07
0.41
0.52
0.04
0.63
0.33
0.16
0.41
0.43
0.14
0.39
0.47
0.10
0.44
0.46
0.13
0.36
0.52
0.07
0.49
0.45
0.08
0.33
0.60
ACTIONS HUMANITAIRES
ALIMENTATION
AMEUBLEMENT DECORATION
APPAREILS MENAGERS
AUDIOVISUEL PHOTO CINEMA
AUTOMOBILE TRANSPORT
BATIMENT TRAVAUX PUBLICS
BOISSONS
CULTURE & LOISIRS
DISTRIBUTION
EDITION
ENERGIE
ENSEIGNEMENT FORMATION
Secteur

B2


SANTE
Topic 0
dcouvrez
cheveux
sommeil
bio
nuit
0 0.005 0.01 0.015 0.02
Topic 1
bion
immunitaires
defense
dfenses
immunit
0 0.02 0.04 0.06 0.08
Topic 2
stress
resist
rsister
rosea
rhodiola
0 0.1 0.2 0.3
Topic 3
peau
oenobiol
soleil
solaire
autobronzant
0 0.05 0.1 0.15
Topic 4
masques
yeux
masque
optone
dispositifs
0 0.05 0.1
Topic 5
microbiosys
ventre
digestif
fort
intestin
0 0.05 0.1 0.15
Topic 6
diarrhe
estomac
intestinal
mdicament
ballonnements
0 0.02 0.04 0.06 0.08
Topic 7
mdicament
risques
exposer
symptmes
parlez
0 0.05 0.1

C1


CULTURE & LOISIRS
Topic 0
gagner
euros
millions
sport
parier
0 0.005 0.01 0.015 0.02
Topic 1
harry
potter
magique
magical
barbie
0 0.02 0.04 0.06 0.08
Topic 2
fitness
park
mois
coach
dpasser
0 0.1 0.2
Topic 3
hatchimals
families
sylvanian
baby
pixies
0 0.05 0.1 0.15 0.2
Topic 4
auto
circuit
tour
course
flash
0 0.02 0.04
Topic 5
cry
babies
larmes
baby
pleurent
0 0.1 0.2 0.3
Topic 6
vip
pets
glam
cheveux
cr
0 0.05 0.1 0.15
Topic 7
concert
paris
france
concerts
jazz
0 0.05 0.1 0.15

C2